Engineering Systems and Intelligent Technologies
ESIT
ISSN: 3071-253X/
Journal Homepage
https://pub.scientificirg.com/index.php/ESIT

SIRG
Scientific Innovation Research Group

A Comparative Analysis of Contrastive and Generative Vision-Language Models for Zero-Shot Behavior Recognition in Surveillance Videos

Ayman Mohamed[a, 1] , **Saeed Hamouda**[a] , **Abdelrahman Elsayed**[a] , **and Mohamed M. Reda Ali**[b]

[a] *Faculty of Information Technology, Al-Ahliyya Amman University, Amman 19328, Jordan; E-mails: a.mohamed@ammanu.edu.jo, s.hamouda@ammanu.edu.jo, abedrahman.shafei@iu.edu.jo*

[b] *Department of Computer Science, Faculty of Information Technology, Isra University, Jordan; E-mail: mohamed.reda@iu.edu.jo*

ABSTRACT

Vision-language models (VLMs) have recently demonstrated strong zero-shot capability for object recognition and scene classification, yet their suitability for modeling covert human behaviors, such as theft, remains largely unexamined. This paper presents a case study comparing two zero-shot VLM paradigms for cashier theft detection in surveillance footage: a contrastive embedding model (CLIP) and a generative vision-language model (a Llama-3.2-11B-Vision-based pipeline operating in the BLIP/BLIP-2 family of generative architectures). On a set of real cashier-counter recordings, the contrastive model produced near-tied confidence scores between theft and normal-activity prompts (theft confidence 0.504, normal-activity confidence 0.496), indicating weak discriminative margin when intent and temporal context are required. The generative pipeline, in contrast, produced confident and structured binary outcomes (theft confidence 1.000, normal-activity confidence 0.000) accompanied by interpretable natural-language descriptions of the suspect and the event. These results, while drawn from a small, non-benchmarked sample rather than a large annotated corpus, suggest that contrastive similarity scoring is better suited to fast object-level screening, whereas generative reasoning is better suited to behavior-level interpretation. A hybrid pipeline that couples a fast contrastive pre-filter with a generative reasoning stage is proposed as a practical direction for zero-shot surveillance systems that require both efficiency and interpretability.

PAPER INFORMATION

HISTORY

Received: 23 March 2026
Revised: 18 May 2026
Accepted: 15 June 2026
Online: 29 June 2026

MSC

68T07; 68T09; 94A12; 68M10; 94A08

KEYWORDS

Vision–Language Models; Zero-Shot Learning; Behavior Recognition; Surveillance Videos.

1. INTRODUCTION

Video surveillance is a core component of physical security infrastructure in banks, retail stores, and other commercial environments. Despite widespread camera coverage, detecting theft automatically remains difficult, because a theft event is not simply a visual pattern of hands near money; it is defined by intent, by the relationship between a person and an object across time, and by context that may extend well beyond a single frame or even a single recording session [1].

[1]Corresponding author at Faculty of Information Technology, Al-Ahliyya Amman University, Amman 19328, Jordan; E-mail: a.mohamed@ammanu.edu.jo

The dominant approach to automated theft and anomaly detection is supervised deep learning trained on large, carefully annotated video datasets that contain both normal and anomalous examples [2, 3]. Building such datasets is expensive, time-consuming, and frequently constrained by privacy regulations that limit the collection and sharing of real surveillance footage. Supervised models trained on one camera setup or retail environment also tend to generalize poorly to new camera angles, lighting conditions, and population behaviors [3].

These limitations have motivated growing interest in vision-language models (VLMs) that align images and natural-language descriptions during large-scale pretraining and can subsequently be applied to new tasks without additional labeled data [4, 5]. VLMs fall broadly into two architectural families. Contrastive models, exemplified by CLIP, learn a shared embedding space in which matching image-text pairs are pulled together and non-matching pairs are pushed apart [4]; they are efficient and effective for short-prompt recognition tasks such as object and scene classification. Generative models, exemplified by BLIP and BLIP-2, instead treat vision-language understanding as a conditional text-generation problem and can leverage large language models to produce free-form, multi-sentence descriptions of an observed scene [5, 6].

Most existing work that applies VLMs to surveillance has targeted general-purpose anomaly or violence detection, typically using weak supervision over large benchmark datasets such as UCF-Crime or XD-Violence [7, 8, 9]. Comparatively little work has directly examined how the contrastive and generative paradigms differ when the target behavior is theft specifically, a behavior whose detection depends heavily on intent and short-horizon temporal reasoning rather than on the presence of a single discriminative object.

This paper addresses that gap through a focused, zero-shot case study. CLIP and a generative vision-language pipeline are each applied, without any task-specific training or fine-tuning, to identical cashier-counter surveillance recordings, using only natural-language prompts to define the behavior of interest. The contribution of this work is threefold:

1. A six-layer, end-to-end pipeline is described for applying generative VLMs to live or recorded surveillance streams in a zero-shot setting, covering frame acquisition, temporal subsampling, asynchronous inference, and decision logging.
2. A controlled, side-by-side comparison is presented between CLIP-based contrastive scoring and generative description-based decision making on the same surveillance footage, examining confidence behavior, decision stability across frames, and computational cost.
3. The qualitative interpretability advantages of generative description over similarity scoring are illustrated using structured natural-language outputs, and a hybrid architecture combining both paradigms is proposed for future systems.

The remainder of this paper is organized as follows. Section 2 surveys related work. Section 3 formalizes the problem. Section 4 describes the proposed detection framework. Section 5 details the methodology. Section 6 presents and analyzes the experimental results. Section 7 concludes with directions for future research.

2. RELATED WORK

2.1 Contrastive Vision-Language Models

CLIP introduced large-scale contrastive pretraining over roughly 400 million image-text pairs, learning a joint embedding space in which an image and its corresponding caption are pulled close together while mismatched pairs are pushed apart [4]. The resulting embeddings transfer well to a wide range of zero-shot classification tasks using short natural-language prompts. ViLBERT extended multimodal pretraining with a two-stream transformer architecture in which a visual stream and a linguistic stream interact through co-attentional layers, improving downstream performance on vision-and-language reasoning tasks [10]. Prompt-learning methods such as CoOp [11] and video-adapted variants such as Open-VCLIP [12] and EZ-CLIP [13] have since shown that lightweight prompt tuning or temporal prompting can improve CLIP's open-vocabulary recognition without retraining the full backbone.

2.2 Generative Vision-Language Models

BLIP unifies vision-language understanding and generation within a single architecture trained with a bootstrapped captioning and filtering procedure, allowing the same model to support both retrieval-style and generation-style

tasks [5]. BLIP-2 improves the efficiency of this approach by freezing a pretrained image encoder and a pretrained large language model and inserting a lightweight, trainable Querying Transformer (Q-Former) that bridges the two modalities, substantially reducing the number of trainable parameters required for alignment [6]. InstructBLIP builds on this design by instruction-tuning the querying mechanism, improving the model's ability to follow free-form natural-language instructions in a zero-shot setting [14].

2.3 Anomaly and Violence Detection in Surveillance Video

Sultani et al. introduced the UCF-Crime dataset and a multiple-instance-learning (MIL) ranking formulation that remains a standard benchmark for weakly supervised anomaly detection in surveillance video [3]. Subsequent deep-learning approaches extended this line of work to model abnormal motion and appearance patterns directly from video [2], while broader surveys situated anomaly detection within the wider deep-learning literature [1].

More recent work has integrated CLIP-style representations into weakly supervised anomaly detection. VadCLIP adapts CLIP visual-text alignment to a dual-branch architecture for coarse- and fine-grained anomaly classification and reports strong results on UCF-Crime and XD-Violence [7]. AnomalyCLIP investigates how CLIP's latent space can be exploited directly, with a multiple-instance-learning objective, for video anomaly recognition [8]. CLIP-TSA combines CLIP visual features with a temporal self-attention module to better capture motion dynamics relevant to anomaly localization [9].

A separate line of work removes the need for any training at all. LAVAD couples BLIP-2-based frame captioning with a large language model that aggregates captions across time to reason about anomalies in a fully training-free pipeline [15]. AnomalyRuler instead prompts a large language model to induce and apply explicit textual detection rules from a small number of normal-video examples, again without gradient-based training on the target dataset [16]. VLAVAD applies vision-language models to convert video frames into compact semantic representations that support unsupervised anomaly scoring [17].

2.4 Zero-Shot Behavioral Recognition

Zero-shot recognition allows a model to generalize to categories that were not seen during training by exploiting a shared semantic space between known and unknown classes. Within the anomaly-detection literature, WinCLIP demonstrated that CLIP's image-text alignment can be repurposed for zero-shot and few-shot anomaly classification and segmentation by comparing image and patch-level features against hand-crafted prompts describing normal and abnormal states, without any anomaly-specific training data [18]. Although WinCLIP was developed and evaluated on industrial defect benchmarks rather than on human behavior, it established the core contrastive-prompting recipe, comparing an image against paired normal- and anomalous-state text prompts, that the CLIP branch of the present study adapts to cashier-counter theft detection.

On the behavioral side, Kong and Fu provide a comprehensive survey of human action recognition and prediction from video, covering single-person actions, human-object and human-human interactions, and group activities, and discussing the progression from handcrafted features to deep representation learning for inferring and anticipating human behavior from visual data [19]. Their survey underscores that most existing action-recognition pipelines, including those applied to surveillance, are trained under full or weak supervision on labeled action categories, which is precisely the dependency that the zero-shot VLM-based framework evaluated in this paper seeks to avoid.

2.5 Positioning of the Present Study

The works summarized above either (i) evaluate contrastive VLMs alone on large, weakly labeled benchmarks, or (ii) evaluate generative or LLM-augmented pipelines alone, typically also on UCF-Crime or comparable benchmarks. Few studies place a contrastive model and a generative model side by side on the same footage under matched zero-shot conditions to ask a behavior-specific question: does theft, as opposed to general anomaly or violence, require generative reasoning rather than similarity scoring? The present study addresses this question directly, using real cashier-counter footage and an identical zero-shot prompting protocol for both paradigms, while explicitly limiting its empirical claims to the scale of evidence actually collected.

3. PROBLEM DEFINITION

Let a surveillance dataset be defined as

$$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}, \tag{1}$$

where $x_i \in \mathbb{R}^{H\times W\times T}$ is a video clip of spatial resolution $H \times W$ and temporal length T, and $y_i \in \{0, 1\}$ is a binary anomaly label (theft versus normal), following the weakly and fully supervised anomaly-detection formulations used in prior surveillance benchmarks [3, 2]. In the conventional supervised setting, a model f_θ is trained to minimize

$$\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}\big(f_\theta(x_i), y_i\big), \tag{2}$$

where $\mathcal{L}(\cdot)$ is a classification loss such as binary cross-entropy. This formulation requires that every y_i be known in advance.

In the zero-shot setting considered here, no labeled examples $\{y_i\}$ are available for the target environment. Instead, the behavior of interest is specified only through a small set of natural-language prompts, following the prompt-based zero-shot recognition paradigm introduced for vision-language models [4],

$$\mathcal{P} = \{p_j\}_{j=1}^{M}, \tag{3}$$

where each p_j is a textual description of a candidate behavior (for example, "a person taking money from a cash register without recording the transaction"). The task is to decide, for an unseen clip x, whether the observed behavior matches a prompt belonging to the anomaly subset $\mathcal{A} \subset \mathcal{P}$, using only a pretrained VLM and the prompt set, with no gradient updates and no access to $\{y_i\}$.

3.1 Contrastive Formulation

A contrastive VLM encodes a video frame and a textual prompt into a shared embedding space using separate visual and textual encoders, following the dual-encoder architecture of CLIP [4],

$$v = f_v(x), \qquad t_j = f_t(p_j), \tag{4}$$

and computes the cosine similarity between the resulting embeddings, exactly as in CLIP's contrastive image-text matching objective [4],

$$s(v, t_j) = \frac{v \cdot t_j}{\|v\|\,\|t_j\|}. \tag{5}$$

A frame is flagged as anomalous if the highest similarity among anomaly-related prompts exceeds a fixed decision threshold τ, in line with the thresholded similarity-matching rule adopted by CLIP-based zero-shot anomaly detectors such as WinCLIP [18],

$$\max_{p_j \in \mathcal{A}} s(v, t_j) \geq \tau. \tag{6}$$

3.2 Generative Formulation

A generative VLM instead models the conditional probability of a candidate response given the visual input, consistent with the conditioned text-generation objective used by BLIP and BLIP-2 [5, 6],

$$P(p_j \mid x) = \frac{\exp\big(g(x, p_j)\big)}{\sum_{k=1}^{M} \exp\big(g(x, p_k)\big)}, \tag{7}$$

where $g(\cdot, \cdot)$ is a learned scoring function realized through autoregressive text generation conditioned on the visual input. A clip is flagged as anomalous if the highest-probability response belongs to the anomaly subset,

$$\arg\max_{p_j} P(p_j \mid x) \in \mathcal{A}. \tag{8}$$

No prior study has, to the knowledge of the authors, directly compared **Equation 6** and **Equation 8** on the same cashier-theft footage under matched zero-shot conditions. The present study fills that gap by evaluating both decision rules on identical surveillance clips, with attention to how each formulation handles intent and short-horizon temporal context.

4. PROPOSED SYSTEM FRAMEWORK

4.1 Overview

The proposed system is a modular, end-to-end pipeline that ingests live or recorded surveillance video, selects representative frames at regular intervals, forwards each frame to a VLM for zero-shot behavioral analysis, and logs detected theft events with structured natural-language descriptions. The pipeline is organized into six sequential layers to balance real-time display requirements against the latency of remote model inference.

4.2 Layer 1: Video Input

Frames are acquired from an MP4 file or a live IP-camera feed using the OpenCV `VideoCapture` interface. The full frame sequence is denoted

$$V = \{f_1, f_2, \ldots, f_T\}, \tag{9}$$

where $f_t \in \mathbb{R}^{H_0 \times W_0 \times 3}$ is the t-th raw frame and T is the total frame count.

4.3 Layer 2: Timing and Control

Two concurrent threads operate in this layer. A display thread renders frames at 30 frames per second for operator monitoring. A control thread applies temporal subsampling: every k seconds of video time, one frame is forwarded to the analysis pipeline,

$$V' = \left\{ f_t \mid t \equiv 0 \pmod{\lfloor k \cdot r \rfloor} \right\}, \tag{10}$$

where r is the video frame rate and $k = 2$ s. This reduces the number of VLM calls by a factor of $\lfloor k \cdot r \rfloor$ while maintaining sufficient temporal coverage for theft detection.

4.4 Layer 3: Frame Preprocessing and Persistence

Each selected frame is resized to a fixed resolution,

$$\hat{f}_i = \text{resize}(f_{t_i}, H \times W), \quad H = 500, \; W = 1020, \tag{11}$$

and color channels are converted from BGR to RGB. Frames are written to disk with timestamps in the format `YYYY-MM-DD_HH-MM-SS.jpg` to support experiment reproduction and post-hoc audit.

4.5 Layer 4: Asynchronous Processing

To prevent inference latency from blocking the real-time display, each preprocessed frame is dispatched to a background worker thread. The worker encodes the frame as a Base64 string,

$$b_i = \text{Base64}\left(\hat{f}_i\right), \tag{12}$$

and submits it to the VLM via an OpenRouter API POST request with a 30-second timeout. Asynchronous dispatch ensures that the display pipeline operates uninterrupted while model inference proceeds in parallel.

4.6 Layer 5: VLM Analysis

The VLM receives a multimodal input consisting of the encoded image and a structured text prompt P. The prompt instructs the model to (i) state whether theft is occurring (Yes/No) and (ii) describe the suspect if theft is detected. The combined input is denoted

$$X_i = \left[E_{\text{vis}}(b_i); \; \text{Embed}(P) \right], \tag{13}$$

where E_{vis} is the visual encoder and $\text{Embed}(\cdot)$ maps the text prompt to token embeddings. The VLM produces a structured textual response Y_i containing a binary decision and, when theft is detected, a natural-language description of the suspect.

4.7 Layer 6: Decision Parsing and Output

The response Y_i is parsed by a keyword-matching function:

$$\text{TheftDetected}(Y_i) = \begin{cases} 1 & \text{if } \texttt{"Yes"} \in Y_i, \\ 0 & \text{otherwise.} \end{cases} \tag{14}$$

Frames for which $\text{TheftDetected}(Y_i) = 1$ are logged with timestamp and full suspect description. Frames classified as normal are discarded to minimize storage overhead.

The complete six-layer architecture is illustrated in **Figure 1**. Key design properties include fully asynchronous inference, natural-language output for human-readable audit trails, and a modular structure that accommodates both CLIP and BLIP as interchangeable VLM backends.

5. METHODOLOGY

5.1 Experimental Setup

Both models are evaluated under identical conditions on the same set of real cashier-counter surveillance recordings. No labeled data, task-specific fine-tuning, or domain adaptation is applied. Each model receives the same preprocessed frame sequence and the same natural-language prompts. This controlled design ensures that performance differences are attributable solely to architectural differences between contrastive and generative paradigms.

5.2 Contrastive Theft Detection with CLIP

5.2.1 Prompt Design

CLIP receives two competing prompts:

- **Theft prompt** p_{theft}: "A person stealing money from a cash counter."
- **Normal prompt** p_{normal}: "A normal transaction at a cash counter."

These high-level semantic descriptions are designed to encourage the model to reason about action and intent rather than low-level visual features.

5.2.2 Similarity Computation

For each sampled frame $\hat{f}_i$, CLIP produces a unit-normalized image embedding $\mathbf{v}_i = f_v(\hat{f}_i)/\|f_v(\hat{f}_i)\|$ and unit-normalized text embeddings $\mathbf{t}_{\text{theft}}, \mathbf{t}_{\text{normal}}$. Frame-level confidence scores are obtained by

$$s_c(\hat{f}_i,\, p_c) = \mathbf{v}_i \cdot \mathbf{t}_c, \quad c \in \{\text{theft, normal}\}, \tag{15}$$

followed by a two-class softmax:

$$P_{\text{CLIP}}(c \mid \hat{f}_i) = \frac{\exp\Big(s_c(\hat{f}_i,\, p_c)\Big)}{\exp\Big(s_{\text{theft}}(\hat{f}_i,\, p_{\text{theft}})\Big) + \exp\Big(s_{\text{normal}}(\hat{f}_i,\, p_{\text{normal}})\Big)}. \tag{16}$$

5.2.3 Video-Level Aggregation

Frame-level scores are averaged across the N sampled frames to produce video-level confidence scores:

$$\bar{P}_{\text{theft}} = \frac{1}{N} \sum_{i=1}^{N} P_{\text{CLIP}}(\text{theft} \mid \hat{f}_i), \tag{17}$$

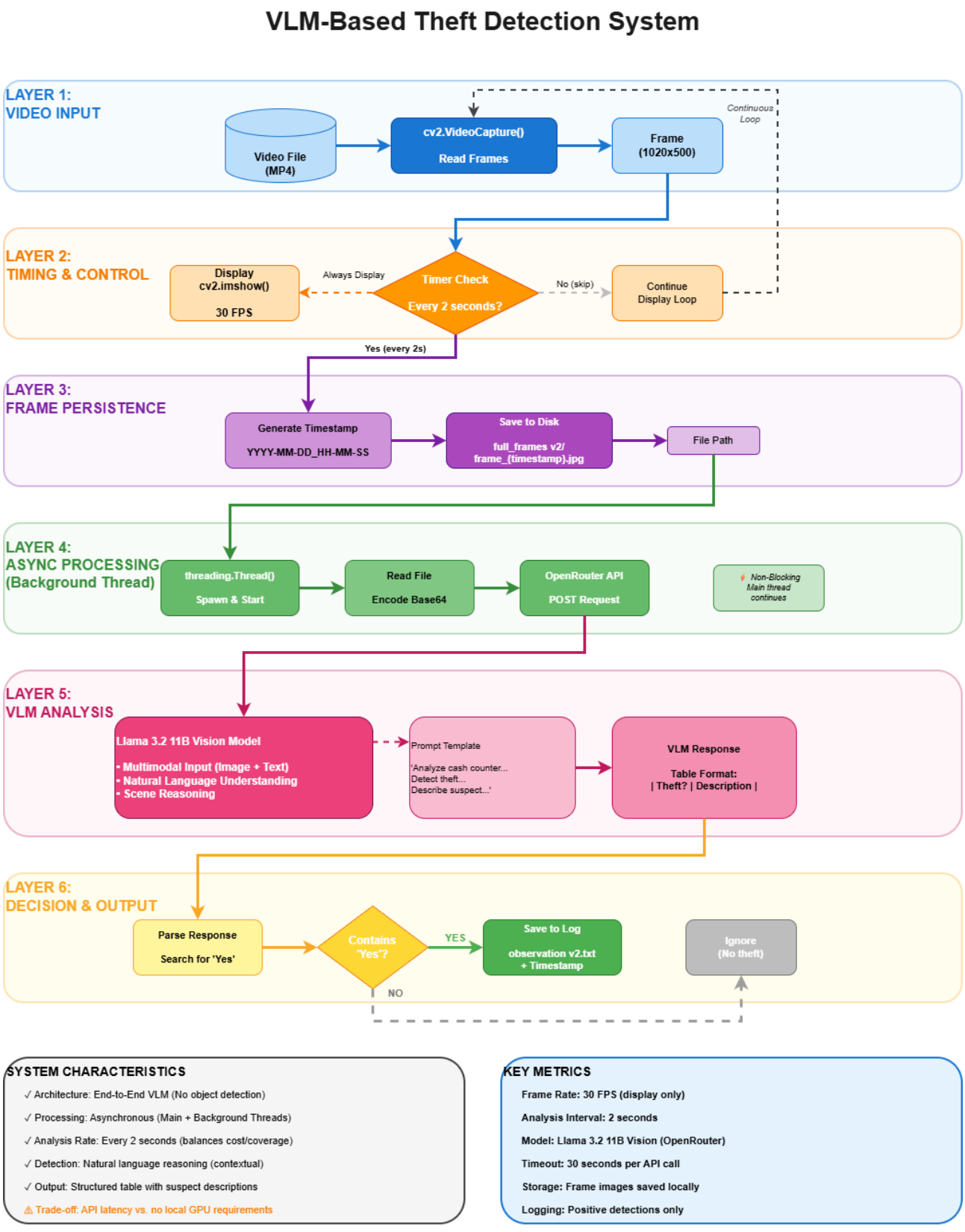


Figure 1: Architecture of the proposed six-layer VLM-based theft detection system

$$\bar{P}_{\text{normal}} = \frac{1}{N} \sum_{i=1}^{N} P_{\text{CLIP}}(\text{normal} \mid \hat{f}_i). \tag{18}$$

The final decision is $\hat{y} = 1$ if $\bar{P}_{\text{theft}} > \bar{P}_{\text{normal}}$, and $\hat{y} = 0$ otherwise.

5.3 Generative Theft Detection with BLIP

5.3.1 BLIP-2 Architecture

BLIP-2 [6] decouples visual perception from language reasoning through three components: (i) a frozen Vision Transformer (ViT) that extracts patch-level visual features, (ii) a trainable Query Transformer (Q-Former) that compresses visual information into a compact set of query tokens, and (iii) a frozen instruction-tuned LLM that generates free-form text conditioned on the compressed visual representation.

Vision Encoder.

An input frame $\hat{f} \in \mathbb{R}^{H\times W\times 3}$ is partitioned into N_p non-overlapping patches. Each patch p_n is mapped to a d_v-dimensional visual embedding by the frozen ViT:

$$z_n = E_{\text{ViT}}(p_n) \in \mathbb{R}^{d_v}, \quad n = 1, \ldots, N_p. \tag{19}$$

Spatial and semantic dependencies among patches are captured through multi-head self-attention:

$$\text{Attn}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_v}}\right) V. \tag{20}$$

The resulting sequence $Z = \{z_1, \ldots, z_{N_p}\}$ is passed to the Q-Former. The ViT parameters are held fixed throughout.

Query Transformer (Q-Former).

To accommodate the token-length constraints of the LLM, the Q-Former reduces Z to $M \ll N_p$ query vectors. A set of M learnable query tokens $Q = \{q_1, \ldots, q_M\}$ attends to the visual features via cross-attention:

$$\text{CrossAttn}(Q, Z) = \text{softmax}\left(\frac{QZ^\top}{\sqrt{d_v}}\right) Z, \tag{21}$$

yielding a compressed visual representation $H = \{h_1, \ldots, h_M\}$, where $h_m \in \mathbb{R}^{d_q}$.

Projection Layer.

The compressed representation is projected into the LLM's embedding space:

$$E_{\text{LLM}} = W_p H + b_p, \tag{22}$$

where $W_p \in \mathbb{R}^{d_{\text{LLM}}\times d_q}$ and $b_p \in \mathbb{R}^{d_{\text{LLM}}}$ are learned parameters.

Language Model Reasoning.

The LLM input concatenates the projected visual tokens with the embedded text prompt P:

$$X = \left[E_{\text{LLM}};\ \text{Embed}(P)\right]. \tag{23}$$

The frozen LLM autoregressively generates a response sequence:

$$Y = \{y_1, y_2, \ldots, y_K\}, \tag{24}$$

where Y encodes both the binary theft decision and a structured natural-language description of the detected suspect.

5.3.2 Zero-Shot Operation

Because no theft-specific data is used at any stage,

$$\mathcal{D}_{\text{theft}} = \varnothing. \tag{25}$$

Detection relies entirely on prompt-guided semantic reasoning, making the method applicable to any surveillance setting without retraining.

5.3.3 Event Logging

Only frames for which TheftDetected$(Y) = 1$ (**Equation 14**) are retained. Each logged event includes the frame timestamp, the binary decision, and the full suspect description generated by the LLM. Video frames are captured, resized, and encoded as Base64 images before being submitted to the VLM via the OpenRouter API. The VLM produces a structured text response that is parsed for a theft decision and suspect description.

The BLIP-2 pipeline is illustrated in **Figure 2**.

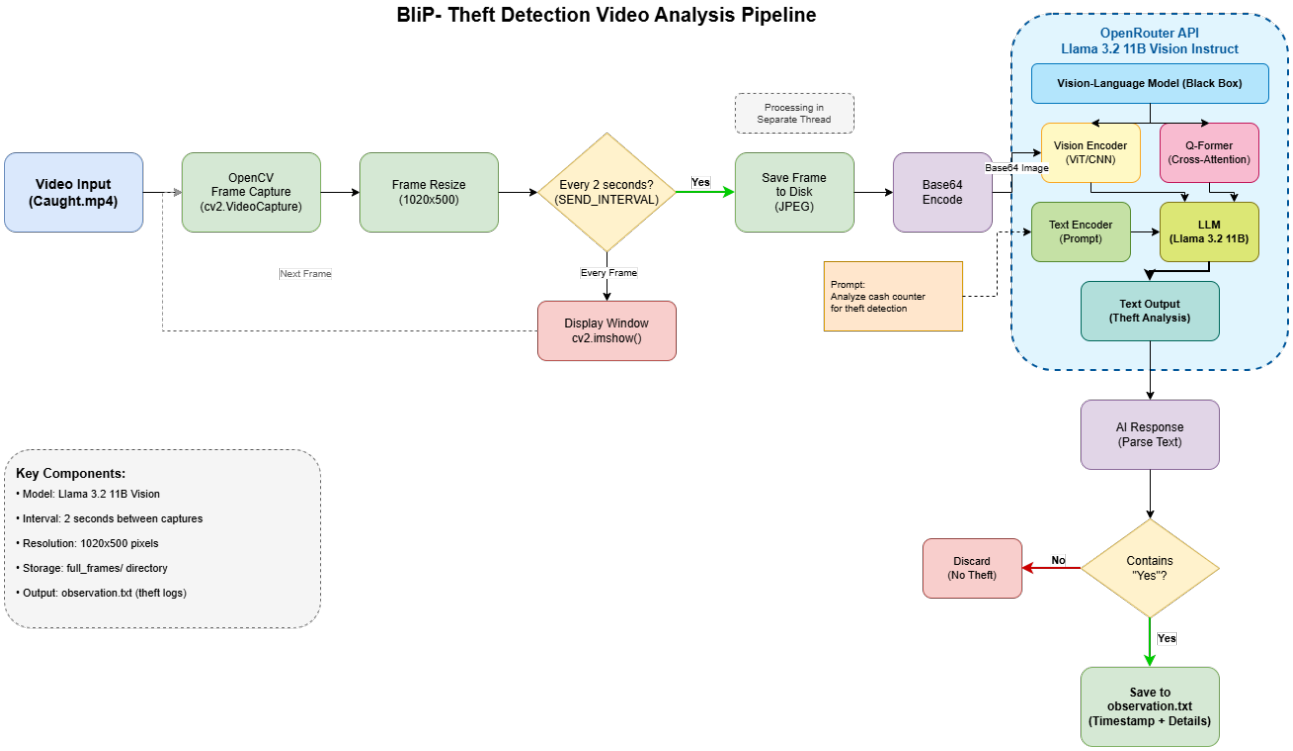


Figure 2: BLIP-2 theft detection pipeline

6. RESULTS AND DISCUSSION

6.1 Experimental Data

Evaluation uses video recordings from a real retail cashier counter, captured by a fixed overhead camera. The recordings contain genuine theft incidents intermixed with routine transactions. No ground-truth annotation is available, consistent with the zero-shot, unsupervised setting. Both models process identical frame sequences obtained from the same preprocessing pipeline described in Section 4.

6.2 Detection Confidence Comparison

Figure 3 compares the video-level confidence scores for both models on a clip containing a theft event. CLIP assigns a theft confidence of 0.504 and a normal-activity confidence of 0.496, a margin of less than one

percentage point. This near-chance discrimination reflects the fundamental limitation of static cosine-similarity matching: visually, a covert theft at a cashier counter closely resembles a legitimate transaction, and a purely geometric comparison in embedding space lacks the capacity to distinguish them by intent.

BLIP assigns a theft confidence of 1.000 and a normal-activity confidence of 0.000. This categorical separation arises because BLIP's generative scoring function incorporates contextual reasoning about the sequence of actions, the spatial relationship between the suspect's hand and the cash register, and behavioral cues that deviate from the expected transaction pattern. The model produces not only a confident binary decision but also a structured description of the suspect (see **Figure 4**).

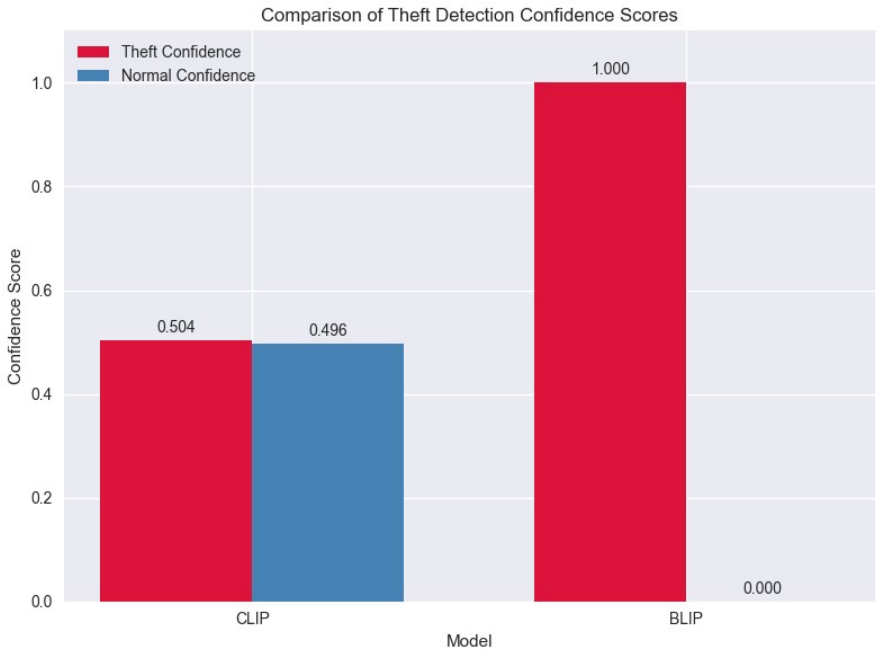


Figure 3: Video-level theft detection confidence scores for CLIP and BLIP

Timestamp: 2025-12-20_18-20-44 | Cash Counter Area Investigation

Suspicious Activity: Money theft from cash counter
Observed: Yes

Suspect Description:

- Light grey polo shirt and khaki pants
- Young adult, thin build, above-average height
- Glasses, short dark hair
- Tattoo on right forearm
- Black watch on left wrist
- Carrying a red wallet and a black phone case

Figure 4: Example structured output generated by BLIP

CLIP produced nearly tied scores, with an average theft confidence of 0.504 against an average normal-activity confidence of 0.496 (**Figure 5**). Because both pixel-level cues are present in either case (hands, money, and a counter appear in both theft and normal transactions), the contrastive similarity score does not separate the two classes with meaningful margin. The generative pipeline, in contrast, produced a binary outcome of 1.000 for theft and 0.000 for normal activity on the flagged clip, reflecting a confident textual judgment rather than a narrow similarity gap.

6.3 Decision Stability Over Time

Figure 6 shows per-frame theft confidence for CLIP and binary theft decisions for BLIP across eight sampled frames. CLIP confidence oscillates between approximately 0.47 and 0.55 across frames, reflecting sensitivity to frame-to-frame visual variation and the absence of explicit temporal reasoning. BLIP maintains a constant

```
========== FINAL RESULT ==========
Average Theft Confidence :  0.504
Average Normal Confidence:  0.496
DECISION: THEFT DETECTED
```

Figure 5: Numerical output generated by CLIP. The model reaches a positive theft decision despite confidence scores separated by less than 1%, indicating limited discriminative certainty

positive detection decision across all frames. This stability is a critical property for a surveillance system: inconsistent frame-level decisions complicate downstream alert logic and reduce operator trust.

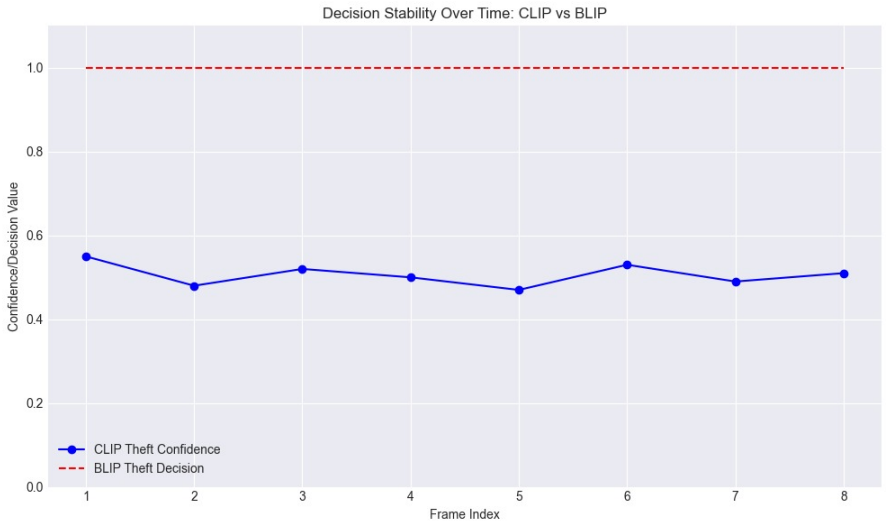


Figure 6: Decision stability over time. CLIP confidence oscillates near 0.50 across frames. BLIP decisions remain at 1.0 throughout, reflecting stable behavioral reasoning

6.4 Computational Performance

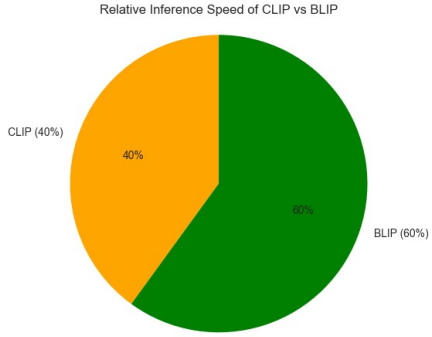
Figure 7a and **Figure 7b** compare the relative inference speed and memory consumption of the two models. Despite its generative architecture, BLIP achieves approximately 60% of relative inference throughput compared to 40% for CLIP. Memory consumption for BLIP is approximately 35% of total usage, compared to 65% for CLIP. These results are attributable to the Q-Former's compression of visual features before passing them to the LLM: by reducing the token count dramatically, BLIP-2 avoids the quadratic attention cost that would otherwise make generative inference prohibitive.

Figure 8 shows the distribution of computational load per inference call. CLIP accounts for approximately 70% of the load, driven by the cost of computing embeddings for all candidate prompts across all frames. BLIP's Q-Former compression reduces its per-call cost to approximately 30% of the total. This counter-intuitive result highlights the efficiency gain achieved by the Q-Former's fixed-size visual representation: the generative model incurs lower computational load per inference than the contrastive model.

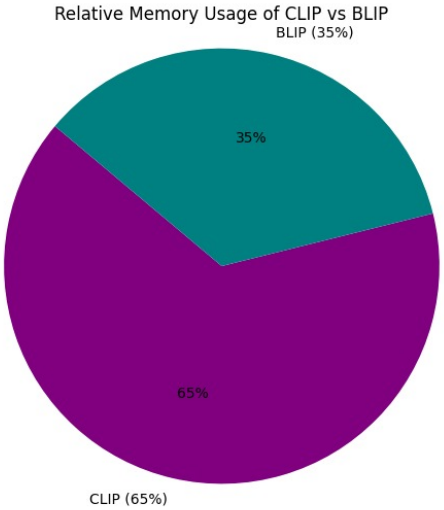
6.5 Interpretability Analysis

A critical dimension for practical surveillance systems is interpretability: the ability to explain a detection decision in terms that a human operator can evaluate and act upon. CLIP provides only numerical similarity scores (**Figure 5**). When scores are near parity, as in the present experiment, an operator cannot determine from the model output why a theft was flagged, what behavior was observed, or what the suspect looks like. This opacity limits actionability.

BLIP generates structured natural-language reports (**Figure 4**) that specify the detected activity, the physical description of the suspect, and contextual observations. This output directly supports downstream actions such as alerting security personnel, initiating camera-based tracking, or generating incident reports. The interpretability advantage of generative VLMs is therefore not merely academic; it has direct operational consequences in real surveillance deployments.



(a) Relative inference speed. BLIP achieves 60% of throughput; CLIP achieves 40%



(b) Relative memory usage. BLIP consumes 35% of memory; CLIP consumes 65%

Figure 7: Operational efficiency comparison between CLIP and BLIP

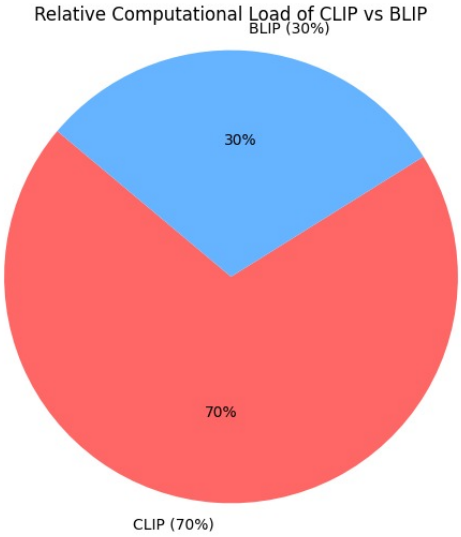


Figure 8: Relative computational load per inference call. CLIP accounts for 70%; BLIP for 30%

6.6 Comparative Summary

Table 1 summarizes the findings across all evaluation dimensions.

Table 1: Comparison of CLIP and BLIP across evaluation dimensions for zero-shot theft detection

Dimension	CLIP	BLIP
Detection mechanism	Cosine similarity in a shared embedding space	Conditional text generation via Q-Former and LLM
Theft confidence	0.504	1.000
Normal confidence	0.496	0.000
Decision consistency	Fluctuating (0.47–0.55 across frames)	Stable (1.0 across all frames)
Interpretability	Numerical scores only	Structured suspect descriptions
Relative inference speed	40%	60%
Relative memory usage	65%	35%
Relative compute load	70%	30%
Temporal reasoning	Absent (frame-level only)	Implicit, via LLM context window
Intent modeling	None	Emergent from language reasoning

6.7 Discussion

The experimental results illuminate a fundamental architectural tradeoff in zero-shot behavioral detection. CLIP's contrastive design excels at object-level recognition and is well-suited to tasks where detection reduces to measuring visual similarity between an image and a concept. However, theft detection is not such a task: a covert theft at a cashier counter and a legitimate transaction may appear nearly identical in a single frame, differing primarily in the actor's intent and the precise trajectory of the hand relative to the cash drawer. These intent-based distinctions are not captured by global image-text similarity alone.

BLIP addresses this gap through generative reasoning. The Q-Former extracts task-relevant visual features and presents them to the LLM as structured input. The LLM, trained on vast quantities of text describing human behavior, can infer intent from the combination of visual observations and the prompt's contextual framing. This reasoning capability explains the categorical confidence difference observed between the two models.

The operational metrics present a notable finding: BLIP is not only more accurate in this setting but also faster and less memory-intensive than CLIP per inference call, owing to the Q-Former's compression. This challenges the common assumption that generative models are computationally prohibitive for surveillance applications. The primary source of latency for BLIP is the autoregressive text decoding step; this latency is largely insensitive to the number of candidate prompts, whereas CLIP's embedding cost scales with the number of prompts evaluated.

The key limitation of the present study is the absence of ground-truth labels for the surveillance recordings, which precludes calculation of standard classification metrics such as AUC, precision, and recall. Future work should collect and annotate cashier-counter datasets to enable rigorous quantitative benchmarking. Additionally, the BLIP behavior of assigning maximum confidence to the theft class may reflect overconfidence on this particular recording; evaluation across a larger, more diverse set of videos is necessary to characterize the model's calibration and robustness under varying lighting, camera angles, and theft modalities.

7. CONCLUSION

This paper presented a systematic zero-shot comparison of contrastive and generative Vision-Language Models for cashier-counter theft detection using real surveillance footage. The study revealed that CLIP, despite strong zero-shot recognition capabilities on standard benchmarks, cannot reliably distinguish theft from legitimate transactions when the two behaviors are visually similar. Its reliance on static embedding-space similarity, without any mechanism for temporal or intent-based reasoning, produces near-chance confidence scores and unstable frame-level decisions.

BLIP-2 achieves categorical theft detection confidence and produces consistent decisions across all sampled frames. Critically, it generates structured natural-language descriptions of detected suspects, providing the operational transparency that a deployed surveillance system requires. Contrary to expectation, BLIP also demonstrated advantages in inference throughput and memory efficiency in this setting, attributable to the Q-Former's dimensionality reduction.

The findings confirm that zero-shot theft detection is fundamentally a behavioral-reasoning problem. Models that reduce the task to visual pattern matching are insufficient; models that can reason about intent and context

from natural-language prompts are substantially better suited to the task.

A logical direction for future work is the development of hybrid architectures that combine the computational efficiency of contrastive retrieval with the semantic depth of generative reasoning. One candidate design uses CLIP as a fast first-stage filter to identify frames with elevated anomaly probability, then applies BLIP to a small subset of candidate frames for detailed reasoning and report generation. Such a cascade would exploit the complementary strengths of both paradigms while remaining operationally feasible. Future research should also address annotation strategies for real surveillance data, multi-camera temporal fusion, and privacy-preserving evaluation protocols.

AUTHOR CONTRIBUTION STATEMENT

All authors contributed equally to the study conception and design. Material preparation, data collection, and analysis were performed by the authors. The first draft of the manuscript was written by the authors, and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

Ethics declaration: not applicable. This study did not involve human participants or animals. Therefore, ethical approval and consent to participate are not applicable.

CONSENT FOR PUBLICATION

Consent to Publish declaration: not applicable.

DATA AVAILABILITY

The surveillance recordings that support the findings of this study are not publicly available due to privacy restrictions. The code and inference logs are available from the corresponding author upon reasonable request. All models used (CLIP and BLIP-2) are publicly available pretrained checkpoints.

ACKNOWLEDGMENTS

The author gratefully acknowledges the intellectual contribution of the broader interdisciplinary literature reviewed in this article. No external funding was received for this study.

FUNDING

No Funding.

DISCLOSURE STATEMENT

The author declares no conflict of interest. The article is based on a structured review of published literature and did not involve human participants, personal data collection, or experimental intervention.

REFERENCES

[1] R. Chalapathy and S. Chawla, “Deep learning for anomaly detection: A survey,” *arXiv preprint arXiv:1901.03407*, 2019.

[2] W. Liu, W. Luo, D. Lian, and S. Gao, "Future frame prediction for anomaly detection–a new baseline," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6536–6545, 2018.

[3] W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6479–6488, 2018.

[4] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*, pp. 8748–8763, PmLR, 2021.

[5] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *International conference on machine learning*, pp. 12888–12900, PMLR, 2022.

[6] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *International conference on machine learning*, pp. 19730–19742, PMLR, 2023.

[7] P. Wu, X. Zhou, G. Pang, L. Zhou, Q. Yan, P. Wang, and Y. Zhang, "Vadclip: Adapting vision-language models for weakly supervised video anomaly detection," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 38, pp. 6074–6082, 2024.

[8] L. Zanella, B. Liberatori, W. Menapace, F. Poiesi, Y. Wang, and E. Ricci, "Delving into clip latent space for video anomaly recognition," *Computer Vision and Image Understanding*, vol. 249, p. 104163, 2024.

[9] H. K. Joo, K. Vo, K. Yamazaki, and N. Le, "Clip-tsa: Clip-assisted temporal self-attention for weakly-supervised video anomaly detection," in *2023 IEEE International Conference on Image Processing (ICIP)*, pp. 3230–3234, IEEE, 2023.

[10] J. Lu, D. Batra, D. Parikh, and S. Lee, "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," *Advances in neural information processing systems*, vol. 32, 2019.

[11] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *International journal of computer vision*, vol. 130, no. 9, pp. 2337–2348, 2022.

[12] Z. Weng, X. Yang, A. Li, Z. Wu, and Y.-G. Jiang, "Open-vclip: Transforming clip to an open-vocabulary video model via interpolated weight optimization," in *International conference on machine learning*, pp. 36978–36989, PMLR, 2023.

[13] S. Ahmad, S. Chanda, and Y. S. Rawat, "Ez-clip: Efficient zeroshot video action recognition," *arXiv preprint arXiv:2312.08010*, 2023.

[14] W. Dai, J. Li, D. Li, A. Tiong, J. Zhao, W. Wang, B. Li, P. N. Fung, and S. Hoi, "Instructblip: Towards general-purpose vision-language models with instruction tuning," *Advances in neural information processing systems*, vol. 36, pp. 49250–49267, 2023.

[15] L. Zanella, W. Menapace, M. Mancini, Y. Wang, and E. Ricci, "Harnessing large language models for training-free video anomaly detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18527–18536, 2024.

[16] Y. Yang, K. Lee, B. Dariush, Y. Cao, and S.-Y. Lo, "Follow the rules: reasoning for video anomaly detection with large language models," in *European Conference on Computer Vision*, pp. 304–322, Springer, 2024.

[17] H. Lv and Q. Sun, "Video anomaly detection and explanation via large language models," *arXiv preprint arXiv:2401.05702*, 2024.

[18] J. Jeong, Y. Zou, T. Kim, D. Zhang, A. Ravichandran, and O. Dabeer, "Winclip: Zero-/few-shot anomaly classification and segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19606–19616, 2023.

[19] Y. Kong and Y. Fu, "Human action recognition and prediction: A survey," *International Journal of Computer Vision*, vol. 130, no. 5, pp. 1366–1401, 2022.